

Molecular Evolution of Glycoside Hydrolase Genes in the Western Corn Rootworm (*Diabrotica virgifera virgifera*)

Seong-il Eyun¹, Haichuan Wang², Yannick Pauchet⁵, Richard H. ffrench-Constant⁵, Andrew K. Benson⁴, Arnubio Valencia-Jiménez^{2,6,7}, Etsuko N. Moriyama^{1,3}, Blair D. Siegfried^{2*}

1 School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, **2** Department of Entomology, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, **3** Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, **4** Food Science and Technology, University of Nebraska-Lincoln, Lincoln, Nebraska, United States of America, **5** Department of Entomology, Max Planck Institute for Chemical Ecology, Jena, Germany, **6** Biosciences, University of Exeter, Penryn, United Kingdom, **7** Departamento de Producción Agropecuaria, Facultad de Ciencias Agropecuarias, Universidad de Caldas, Manizales, Colombia

Abstract

Cellulose is an important nutritional resource for a number of insect herbivores. Digestion of cellulose and other polysaccharides in plant-based diets requires several types of enzymes including a number of glycoside hydrolase (GH) families. In a previous study, we showed that a single GH45 gene is present in the midgut tissue of the western corn rootworm, *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae). However, the presence of multiple enzymes was also suggested by the lack of a significant biological response when the expression of the gene was silenced by RNA interference. In order to clarify the repertoire of cellulose-degrading enzymes and related GH family proteins in *D. v. virgifera*, we performed next-generation sequencing and assembled transcriptomes from the tissue of three different developmental stages (eggs, neonates, and third instar larvae). Results of this study revealed the presence of seventy-eight genes that potentially encode GH enzymes belonging to eight families (GH45, GH48, GH28, GH16, GH31, GH27, GH5, and GH1). The numbers of GH45 and GH28 genes identified in *D. v. virgifera* are among the largest in insects where these genes have been identified. Three GH family genes (GH45, GH48, and GH28) are found almost exclusively in two coleopteran superfamilies (Chrysomeloidea and Curculionoidea) among insects, indicating the possibility of their acquisitions by horizontal gene transfer rather than simple vertical transmission from ancestral lineages of insects. Acquisition of GH genes by horizontal gene transfers and subsequent lineage-specific GH gene expansion appear to have played important roles for phytophagous beetles in specializing on particular groups of host plants and in the case of *D. v. virgifera*, its close association with maize.

Citation: Eyun S-i, Wang H, Pauchet Y, ffrench-Constant RH, Benson AK, et al. (2014) Molecular Evolution of Glycoside Hydrolase Genes in the Western Corn Rootworm (*Diabrotica virgifera virgifera*). PLoS ONE 9(4): e94052. doi:10.1371/journal.pone.0094052

Editor: Omprakash Mittapalli, The Ohio State University/OARDC, United States of America

Received November 21, 2013; **Accepted** March 11, 2014; **Published** April 9, 2014

Copyright: © 2014 Eyun et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Consortium for Plant Biotechnology Research (CPBR Agreement GO12026-333) and Pioneer Hi-Bred International. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Funding for this research was provided by the Consortium for Plant Biotechnology Research (Agreement GO12026-333) with matching support from Pioneer Hi-Bred International. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* E-mail: bsiegfried1@unl.edu

Introduction

Cellulose, which is mostly synthesized by terrestrial plants and marine algae, is the most abundant organic compound on Earth. It is a simple carbohydrate polymer, consisting of repeating glucose units linked by β -1,4-glycosidic bonds. It is comprised of nanometer-thick crystalline microfibrils and highly resistant to enzymatic hydrolysis [1]. Cellulolytic fungi and bacteria have developed complex cellulase systems that efficiently hydrolyze cellulose [2]. These cellulase systems play important roles in a wide range of processes ranging from biosphere maintenance (carbon recycling) to the generation of potentially sustainable energy sources such as glucose, ethanol, hydrogen, and methane [1,3–5].

For many herbivorous, detritivorous, as well as omnivorous insects, cellulose comprises a major nutritional resource. However, endogenous cellulases were long thought to be absent in metazoans including insects. It had been wildly accepted that cellulose digestion in insects was mediated by gut-associated microbes such as mixtures of bacteria and protozoa under

anaerobic conditions [6–8]. However, since the first endogenous cellulase gene was identified in the termite *Reticulitermes speratus* [9], many studies now account for the endogenous origin of cellulases in nematodes, insects, and some other invertebrates [10–12]. In the termite systems, where the metazoan cellulose digestion is most extensively studied, a dual (independent) or synergistic collaboration system among host and symbiont-mediated cellulases has been proposed [13–18]. However, understanding of the exact roles of the host and symbiotic microbiota in the complex cellulose degradation process is still emerging.

Cellulase is a general term for cellulytic enzymes including three classes of hydrolytic enzymes: endoglucanases (EC 3.2.1.4), exoglucanases (cellobiohydrolases: EC 3.2.1.74 and 3.2.1.91), and β -glucosidases (cellobioses: EC 3.2.1.21). Plant cell wall digestion also requires other enzymes including pectinases and hemicellulases. All these enzymes are grouped into glycoside hydrolase (GH; EC 3.2.1.-) (also known as glycosidase or glycosyl hydrolase) families according to their amino-acid sequence similarities and their folding patterns based on the Carbohy-

drate-Active enZymes Database (CAZy, <http://www.cazy.org>) [19]. β -glucosidases (GH family 1) are found universally in all domains of organisms [20]. While insects lack endogenous exoglucanases ([11] but see, e.g., [21]), genes encoding endoglucanases and other GH family enzymes have been identified from a number of phytophagous coleopterans belonging to the superfamilies Chrysomeloidea, which includes long-horned beetles and leaf beetles, and Curculionoidea (weevils) [11,22–24]. A β -1,4-endoglucanase gene belonging to the GH family 9 was also isolated and characterized from the red flour beetle *Tribolium castaneum* (Coleoptera: Tenebrionidae) [25].

Recently, we have cloned and characterized a novel β -1,4-endoglucanase gene (*DvvENGaseI*, JQ755253) belonging to the GH family 45 from the western corn rootworm *Diabrotica virgifera virgifera* (Coleoptera: Chrysomelidae), an important insect pest of maize (*Zea mays* L.) in the United States [26,27]. We showed that suppression of *DvvENGaseI* expression by RNA interference (RNAi) resulted in only slight developmental delays suggesting that this gene might be a part of the larger system of cellulose degrading enzymes [26]. The goal of this study is focused on the exploration of genetic diversity among GH family genes in *D. v. virgifera*, especially focusing on its larval stages. In order to identify the diversity of GH family genes encoding plant cell wall degrading and related enzymes expressed in *D. v. virgifera* larvae, we sequenced the transcriptomes covering three different developmental stages (eggs, neonates, and midgut from third instar larvae) using next-generation technologies. We identified eight types of GH family genes that encode β -1,4-endoglucanases (GH45, GH48, and GH5) as well as a pectinase (GH28), an endo-1,3- β -glucanase (GH16), an α -galactosidase (GH27), an α -glucosidase (GH31), and a β -glucosidase (GH1). We found large numbers of GH45 and GH28 genes from the *D. v. virgifera* transcriptomes, among the largest so far known from coleopteran species studied. Our analyses also suggested multiple horizontal transfer events during the evolution of GH45, GH48, and GH28 genes from bacteria or fungi to the common ancestor of chrysomelid and curculionid beetles as well as to other herbivorous insects. Acquisition and subsequent expansion of GH gene copies in phytophagous beetle lineages may have been adaptive and have played important roles for their specialization in feeding on particular host plants.

Results and Discussion

Sequencing and *de novo* Assembly of *D. v. virgifera* Transcriptomes

Using Illumina paired-end as well as 454 Titanium sequencing technologies, in total ~700 gigabases were sequenced from cDNA prepared from eggs (15,162,017 Illumina paired-end reads after filtering), neonates (721,697,288 Illumina paired-end reads after filtering), and midguts of third instar larvae (44,852,488 Illumina paired-end reads and 415,742 Roche 454 reads, both after filtering) (see Table S1 for details). *De novo* transcriptome assembly was performed using Trinity [28] for each of three samples as well as for the pooled dataset (see Materials and Methods and Tables S1, S2, and S3 for the comparative analysis of assembly programs and other details). The *D. v. virgifera* transcriptome assembled from the pooled dataset included 163,871 contigs (the average length: 914 bp) (Table 1).

Identification of GH Family Genes from *D. v. virgifera* Transcriptomes

A total of seventy eight potential genes belonging to eight GH families were identified from our *D. v. virgifera* transcriptome. In

Figure 1, numbers of the genes for these GH families found in *D. v. virgifera* are compared with those found in other coleopteran species. While the enzymes encoded by GH45, GH48, and GH5 family genes are known to have β -1,4-endoglucanase (EC. 3.2.1.4) activity, GH28 genes encode a pectolytic enzyme, polygalacturonase (EC 3.2.1.15) [19]. GH16 family genes encode a laminarinase, β -1,3-glucanase (EC 3.2.1.39). We also found genes encoding GH27 (α -galactosidase, EC 3.2.1.22), GH31 (α -glucosidase, EC 3.2.1.20), and GH1 (β -glucosidase, EC 3.2.1.21) families.

GH45 Family

Eleven GH45 family gene candidates were identified from the *D. v. virgifera* transcriptome with ten of them covering the entire coding regions (615–741 bp or 204–246 amino acids (AA); Figure S1A). The partial sequence (GH45-6) was also confirmed in the draft *D. v. virgifera* genome. Four of them (GH45-1, GH45-4, GH45-7, and GH45-10) were highly expressed (>100 reads per kilobase of per million mapped reads or RPKM in the neonate and third-instar larval midgut samples but not expressed in the egg samples (Table S4). We have previously identified GH45-7 as *DvvENGaseI* (JQ755253) [26]. This gene exhibits the highest expression among the eleven GH45 family genes and also the highest among all GH genes identified in the present study (Table S4). Note that its gene and protein expressions in *D. v. virgifera* larvae were also confirmed in our previous study [26].

GH45 family genes have been described from a number of coleopteran species belonging to the suborder Polyphaga (e.g., [23,29–31]). Similarity searches against the NCBI (National Center for Biotechnology and Information) non-redundant (NR) protein database as well as ten complete insect genomes confirmed that within insects, GH45 family genes are found only in two polyphagan coleopteran superfamilies, Chrysomeloidea and Curculionoidea. As shown in Figure 1, multiple GH45 genes have been identified in some species. *D. v. virgifera* has the largest known number of GH45 genes (11 genes) among coleopteran species, and probably among any known invertebrates where this gene exists.

In addition to these coleopteran sequences, a sequence similar to the GH45 family has been identified from the Antarctic springtail *Cryptopygus antarcticus* [12], which belongs to one of the basal hexapodan orders, Collembola [32]. Another sequence similar to the GH45 family was reported from the water bear *Hypsibius dujardini* (phylum Tardigrada, a sister group of arthropods) [33]. GH45 family genes have also been reported among various metazoans, from protists symbiotic to wood-feeding termites and a cockroach [34,35] to plant-parasitic nematodes and mollusks [10,36–38]. In order to understand the evolutionary process that has led to the diversity of coleopteran GH45 family genes, a maximum-likelihood phylogeny was reconstructed including GH45 family proteins from eleven coleopteran species as well as other metazoans mentioned above, fungi, and bacteria (Figure 2). Our phylogenetic analysis suggests that all coleopteran GH45 family genes are monophyletic although the support was weak ($\leq 66\%$ bootstrap supports). Based on currently available sequences, several species-specific gene duplications were found in coleopteran species (shown with blue branches in Figure 2). While all bacterial GH45 family proteins, except for sequences from *Myxococcus stipitatus* and uncultured bacterium (ADV57513.1), formed a monophyletic group, relationships among fungal and metazoan sequences were unresolved. Although the exact timings are not clear, multiple horizontal gene transfer (HGT) events are likely to have involved in the evolution of metazoan GH45 family genes.

As shown in [22], except for a clade of curculionid proteins (Group 1 in Figure 2), the putative catalytic nucleophile and

Table 1. Summary of the *D. v. virgifera* transcriptome assembly using the pooled dataset.

Samples	Egg, neonate, and third-instar larval midgut
Number of paired-end reads before filtering	$1,462.2 \times 10^6$ ($144,690 \times 10^6$ bp)
Number of paired-end reads after filtering	781.7×10^6 ($77,393 \times 10^6$ bp)
Assembly program used	Trinity (2013-02-25)
Total number of contigs	163,871
Average contig length (range)	914 bp (201–31,064 bp)
N50 length	1,396 bp

doi:10.1371/journal.pone.0094052.t001

proton donor positions of GH45 proteins are highly conserved with Asp (Gln is found in Group 1 proteins). This is also the case for all but one *D. v. virgifera* GH45 proteins (Val is found in GH45-9; Figure S1A). Other exceptional cases include: Asn in *H. dujardini* (Tardigrada), Thr in *Leptosphaeria maculans* (fungus), Ser in *Alternaria alternate* (fungus), and Glu in *Myxococcus stipitatus* (bacterium). In

addition to possessing the conserved catalytic residues, the majority of GH45 genes identified in *D. v. virgifera* showed significant expression levels in neonate and third-larval midgut transcriptomes (Table S4). These proteins probably share similar functions and help explain our previous results with RNAi-

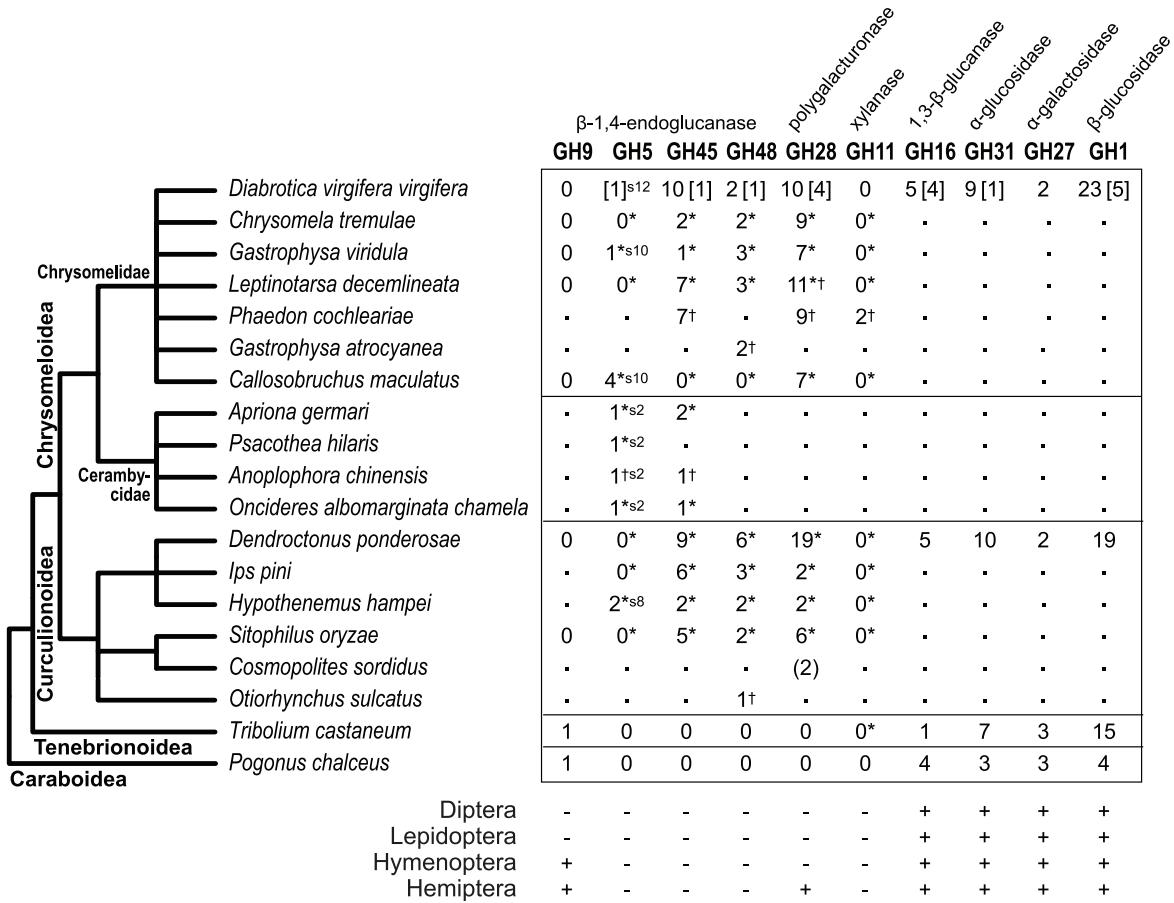


Figure 1. Distribution of glycoside hydrolase family genes among polyphagan coleopterans. Numbers for GH5, GH45, GH48, GH28, and GH11 genes are taken from [22] (marked with *). Exceptions are for *D. v. virgifera* (this study; numbers in square brackets are for partial sequences), *P. cochleariae* GH45, GH28 [83], and GH11 [24], *G. atrocyanea* GH48 [39], *D. ponderosae* [45], *C. sordidus* (preliminary results from transcriptomes are shown in parentheses; A. Valencia-Jiménez, personal communication), *O. sulcatus* GH48 (CAH25542.1), *T. castaneum* [25,70], and *P. chalceus* (this study, searched from the transcriptome [69]). Numbers with † indicate that they are based on the search results from the NCBI NR database or from literatures. Since neither genomes nor transcriptomes are available for these species, the actual numbers of their GH family genes are not known. For GH5 genes, their subfamilies are indicated with 's' followed by the number (e.g., s2 for subfamily 2). Accession numbers for all coleopteran GH genes included in this study are found in Table S5. The taxonomical relationship is based on [68]. ‘-’ not determined. For other insect groups, only existence (+) or absence (-) is shown.

doi:10.1371/journal.pone.0094052.g001

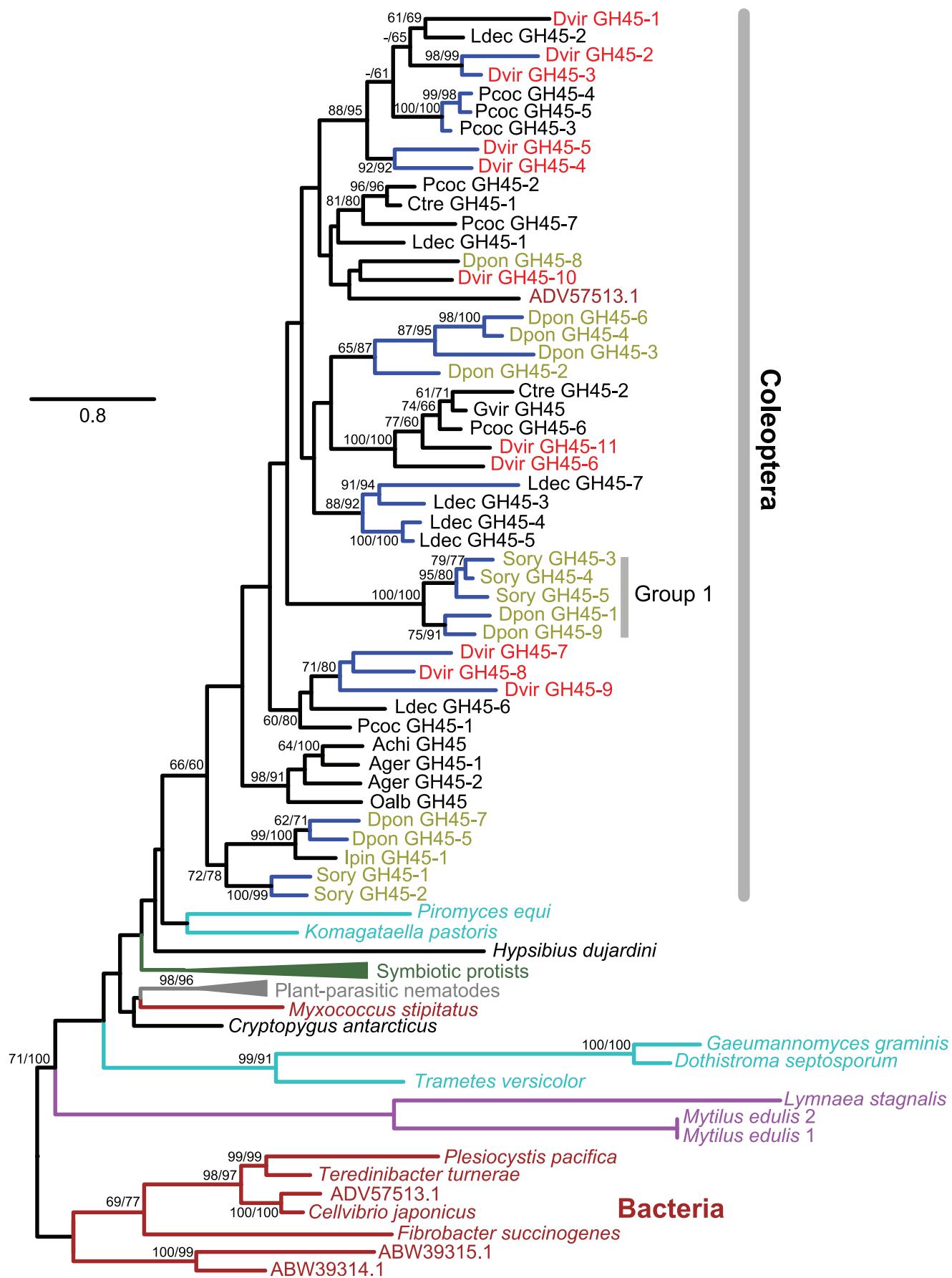


Figure 2. The maximum-likelihood phylogeny of GH45 proteins. Forty seven GH45 protein sequences from eleven coleopteran species are included. Their species name abbreviations are found in Table S5. Labels for the coleopteran species belonging to the superfamily Curculionoidea are olive-colored and all other coleopteran sequences colored in black belong to the superfamily Chrysomeloidea. *D. v. virgifera* sequences are shown in red. Other sequences include: two mollusks (purple), *Cryptopygus antarcticus* (Collembola, black), *Hypsibius dujardini* (Tardigrada, black), 24 termite-symbiotic protists (dark green), 10 plant-parasitic nematodes (all are from *Bursaphelenchus xylophilus*, grey), representative fungi (chosen from 138 sequences, cyan), and representative bacteria (chosen from 18 sequences, brown). Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Supporting values are shown only when higher than 60%. Blue-colored branches indicate the species-specific gene duplications (based on currently available sequences) within a cluster supported by higher than 70% of bootstrap values. The scale bar represents the number of amino acid substitutions per site. See Figure S2 for more details.

doi:10.1371/journal.pone.0094052.g002

suppression of single GH45 gene expression not drastically affecting the *D. v. virgifera* larval development.

GH48 Family

We identified three GH48 family gene candidates from *D. v. virgifera*: two complete (1,926 bp, 641 aa) and one partial (374 bp, 124 aa) (Figure S1B). The partial sequence (GH48-2) was confirmed in the draft *D. v. virgifera* genome. Similar to GH45 family genes, GH48 family genes have been identified from many polyphagan coleopterans especially from the two superfamilies (Chrysomeloidea and Curculionoidea) [22,39,40] (Figure 1). Consistent with the results shown in [22], the number of GH48 genes found in coleopterans was in general smaller than those of GH45 and GH28 family genes.

Two GH48 family genes (active phase-associated proteins, APAP I and II; shown as Gatr GH48-1 and -2 in Figure 3) were isolated from a leaf beetle *Gastrophysa atrocyanea* [39]. While neither glucanase nor cellobiohydrolase activity was detected with *G. atrocyanea* GH48-1, it exhibited chitinase activity. *G. atrocyanea* GH48-1 was shown to be necessary for diapause termination in adults [39]. Based on our phylogenetic analysis, *G. atrocyanea* GH48-1 was found to be closer to *D. v. virgifera* GH48-2 (Figure 3). However, only a fragment has been identified from the *D. v. virgifera* GH48-2 and its expression was not confirmed from our egg and larval samples (Table S4). While *D. v. virgifera* GH48-1 also had very low expression, GH48-3 was found to be expressed more in larvae than in eggs.

GH48 is one of the most common GH family genes in bacteria [41]. Apart from their presence in bacteria and in coleopterans, this family has been reported from three fungal species (*Neocallimastix patriciarum*, *Piromyces equi*, and *Piromyces* sp.). None of the ten insect genomes we examined had GH48 family genes. This disparate and limited distribution of GH48 family genes in two related coleopteran superfamilies and in three fungal species but not in any other eukaryotes, clearly indicates at least two independent HGT events: one from bacteria to the ancestral coleopteran lineage before the divergence of the two coleopteran superfamilies and the other from bacteria to the ancestral lineage before the divergence of the three fungal species. The three fungal GH48 sequences belong to the family Neocallimastigaceae (phylum Neocallimastigmota). These fungi are isolated in the digestive tracts of ruminant and non-ruminant mammals and herbivorous reptiles [42]. Although our similarity search and phylogenetic analysis did not show a clear relationship with any known bacterial species, rumen fungi have been reported to obtain catalytic enzymes from bacterial sources by HGT events. For example, GH5 (endoglucanase, EC 3.2.1.4) and GH11 (xylanase, EC 3.2.1.8) family genes found in *Orpinomyces joyonii* and *Orpinomyces* sp. (phylum Neocallimastigmota) are considered to be bacterial origin [43]. GH5 family genes in a rumen fungus, *Neocallimastix patriciarum*, have also been suggested to have originated from bacteria (*Streptococcus equinus* and *Ruminococcus albus*) [44].

GH28 Family

GH28 family genes encode polygalacturonase (pectinase, EC 3.2.1.15). Ten intact (average 1087 bp, 361 aa) and four partial GH28 candidate sequences were identified in the *D. v. virgifera* transcriptome (Figure S1C). Gene expression, especially in larvae, was confirmed from the majority of the eleven intact candidates (Table S4). Although the expression of the three partial sequences (GH28-8, 10, and 14) was either very low or confirmed neither in eggs nor in larvae, their partial sequences were found in the draft genome. Multiple copies of GH28 family genes have been found in a number of coleopteran species belonging to its two superfamilies (Chrysomeloidea and Curculionoidea) [22,29]. The largest number (19 functional genes) was found in the mountain pine beetle (*Dendroctonus ponderosae*) [40,45]. *D. v. virgifera* has the second largest number, 10 (and 4 partial sequences), of GH28 family genes (Figure 1). Our phylogenetic analysis based on currently available sequences confirmed many species-specific duplications of GH28 family genes in coleopterans (Figure 4, blue branches).

Consistent with what was indicated by Pauchet *et al.* [22], our phylogenetic analysis showed that GH28 family genes can be divided into two clades. GH28 enzymes from *Callosobruchus maculatus* (bean beetle) form a subgroup (B, Figure 4) and are more closely related to bacterial GH28 enzymes (all Gram-negative bacteria) (>83% bootstrap supports), while all other beetle GH28 enzymes are more closely related to fungal and plant bug (Hemiptera) enzymes. Although two plant bug species (*Lygus hesperus* and *Lygus lineolaris*, Hemiptera) were reported to have multiple GH28 family genes [46,47], we failed to identify GH28 candidate sequences in the ten insect genomes including two from hemipterans *Rhodnius prolixus* (a blood-sucking bug) and *Acyrthosiphon pisum* (pea aphid). Among insects, except for the two plant bug species, GH28 family genes were found only in two coleopteran superfamilies (Chrysomeloidea and Curculionoidea). These insect GH28 family genes except for those of *C. maculatus* are phylogenetically nested within a fungal GH28 cluster (Figure 4). Therefore, GH28 genes currently found in coleopterans and plant bugs were most likely acquired by three independent HGT events: from Gram-negative bacteria to *C. maculatus*, from a fungus to a hemiptera, and from a fungus to an ancestral coleopteran before the divergence of the two superfamilies.

GH16 Family

We identified nine GH16 family genes, which encode β-1,3-glucanases, in the *D. v. virgifera* transcriptome: five full-length (average 1124 bp, 374 aa) and four partial coding sequences (Figure S3A). Their expressions were identified both in larval and egg samples (Table S4). The most significantly highly expressed gene (GH16-1) showed larval specific expression. GH16 family genes are widely found in insects (e.g., [48–50]). Similarity searches further confirmed a wide distribution of GH16 family genes within metazoa including insects, mollusks (e.g., [51]), sea urchins (e.g., *Strongylocentrotus purpuratus*), as well as basal chordates (e.g., *Ciona intestinalis*) but not in vertebrates. It was also found widely in fungi

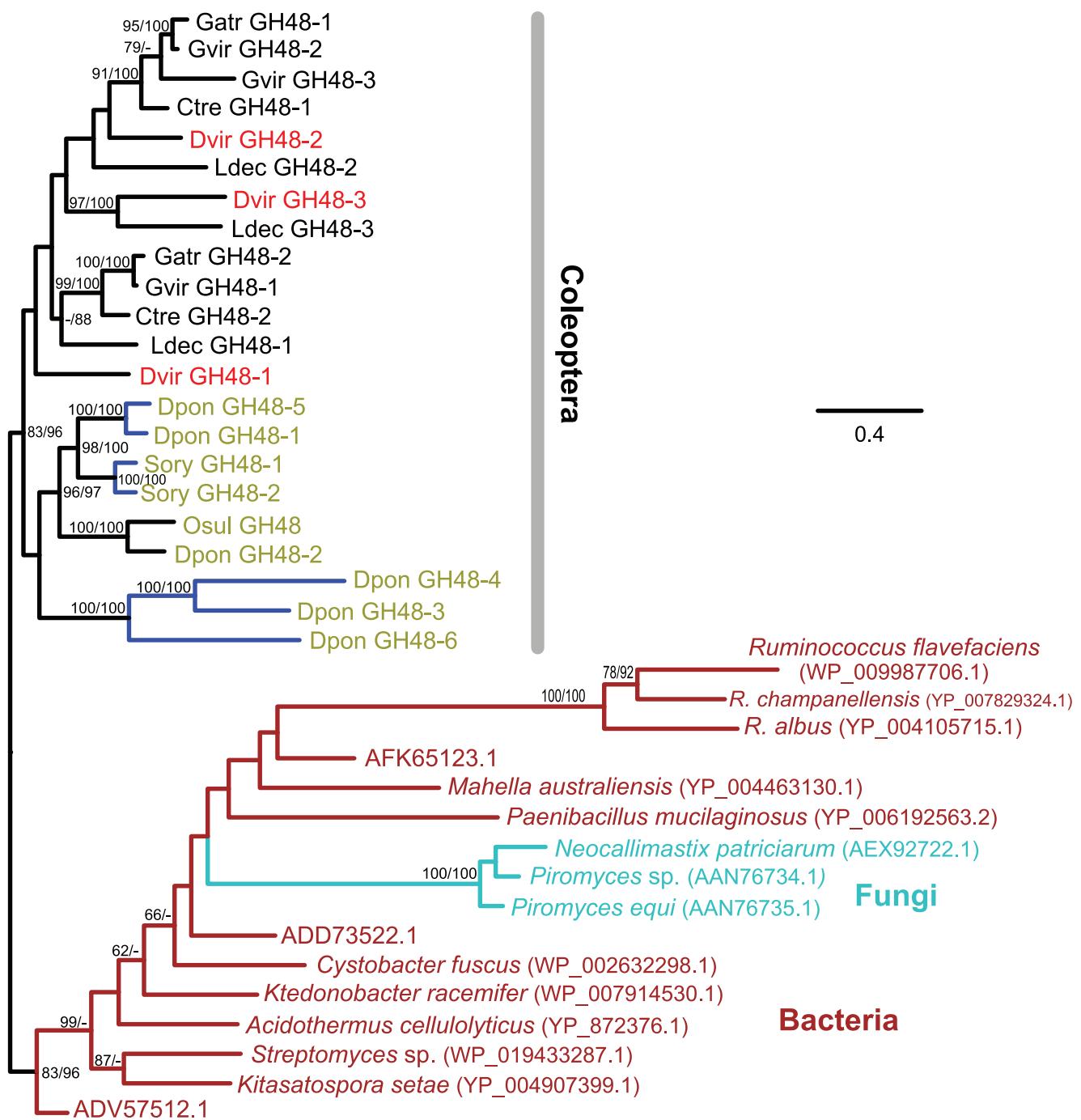


Figure 3. The maximum-likelihood phylogeny of GH48 proteins. Twenty two GH48 protein sequences from seven coleopteran species are included. Their species name abbreviations are found in Table S5. Labels for the coleopteran species belonging to the superfamily Curculionoidea are olive-colored and all other coleopteran sequences colored in black belong to the superfamily Chrysomeloidea. *D. v. virgifera* sequences are shown in red. Other sequences include: representative bacteria (chosen from 653 sequences, brown) and 3 fungi (shown in cyan). Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Supporting values are shown only when higher than 60%. Blue-colored branches indicate the species-specific gene duplications (based on currently available sequences) within a cluster supported by higher than 70% of bootstrap values. The scale bar represents the number of amino acid substitutions per site.
doi:10.1371/journal.pone.0094052.g003

and bacteria. An ortholog identified in the Antarctic springtail *C. antarcticus* (CaLam) is believed to have originated from bacteria by HGT [52].

Figure 5 shows the phylogeny of GH16 family protein sequences from four coleopteran species (*T. castaneum*, *Tenebrio*

molitor, *D. ponderosae*, and *D. v. virgifera*) and other insects as well as some other metazoans, fungi, and bacteria. As reported previously, insects have a group of pattern recognition proteins that are originated from a duplicated copy of GH16 family genes [48–50]. They are called Gram-negative bacteria-binding proteins (GNBPs)

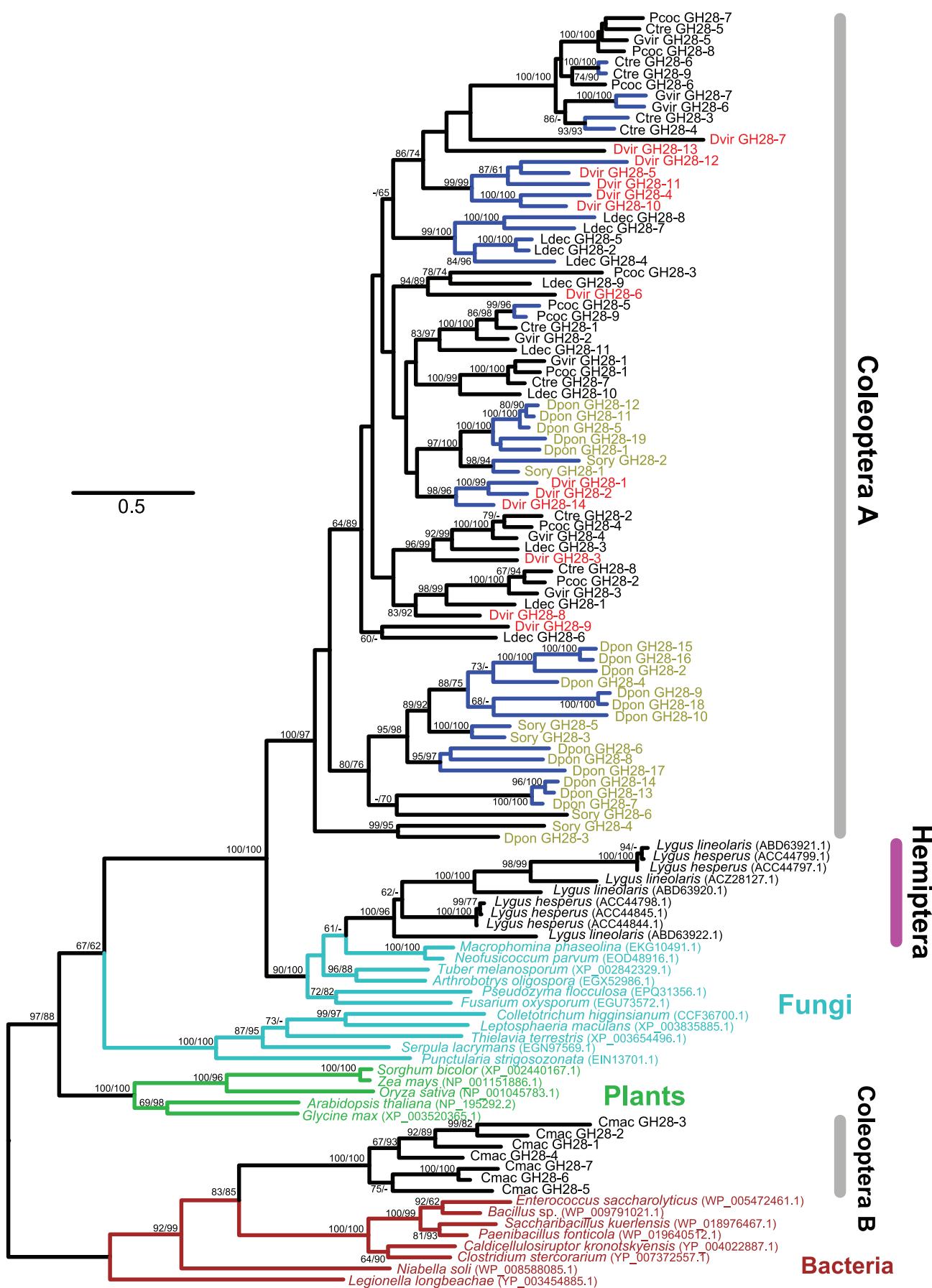


Figure 4. The maximum-likelihood phylogeny of GH28 proteins. Eighty four GH28 protein sequences from eight coleopteran species are included. Their species abbreviations are found in Table S5. Labels for the coleopteran species belonging to the superfamily Curculionoidea are olive-colored and all other coleopteran sequences colored in black belong to the superfamily Chrysomeloidea. *D. v. virgifera* sequences are shown in red. Other sequences include: plant bugs (*Lygus hesperus* and *Lygus lineolaris*), representative fungi (chosen from 651 sequences, cyan), representative bacteria (chosen from 42 sequences, brown), and representative plants (chosen from 491 sequences, green). Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Supporting values are shown only when higher than 60%. Blue-colored branches indicate the species-specific gene duplications (based on currently available sequences) within a cluster supported by higher than 70% of bootstrap values. The scale bar represents the number of amino acid substitutions per site.

doi:10.1371/journal.pone.0094052.g004

or β -1,3-glucan recognition proteins (β GRPs) and are involved in innate immune recognition [53–57]. These proteins (indicated as “GNBP” in Figure 5) have lost their original GH16 enzymatic activity [56,58] and their active sites are not conserved (Figure S3B; also see [48,49]). These proteins, however, contain a unique conserved β -1,3-glucan binding domain in their N-terminal regions [57,59,60]. Both types of genes, GH16 family genes with conserved active sites as well as GNBP genes that contain the N-terminal domain but no conserved active sites, have been identified from the coleopteran species examined so far (*T. castaneum*, *T. molitor*, *D. ponderosae*). Both types of genes were also identified in our *D. v. virgifera* transcriptome. For the nine GH16 family gene candidates, the active site regions show highly conserved patterns including two Glu residues (Figure S3A). We also identified three potential GNBP genes from *D. v. virgifera*. Consistent with GNBP found in other insects, their active sites are not conserved with Glu’s (Figure S3B). All but one GNBP gene candidates were weakly expressed both in eggs and in larvae (neonates and third-instar larval midguts) (Table S4), which is consistent with the pattern found with *Drosophila* GNBP genes [53]. One gene (GNBP-3) does not have the conserved N-terminal domain (Figure S3C); this gene may not function as a GNBP. No expression was detected from this gene in larval samples.

The β -1,3-glucanase activity has been confirmed with GH16 enzymes identified from midguts of several insects including *T. molitor* (TLam) [61], *Spodoptera frugiperda* (Lepidoptera) (SLam) [49], *Helicoverpa armigera* (Lepidoptera) [48], *Abracris flavolineata* (Orthoptera) [62], *Periplaneta americana* (Blattodea) [63], as well as termites (Isoptera) [58]. These insect, except for lepidopteran, enzymes can lyse *Saccharomyces cerevisiae* cells. For detritivorous insects such as *T. molitor*, *P. americana*, and termites, their abilities of digesting fungal cell walls may play roles in antifungal protection as well as in nutrient acquisition. *T. molitor*’s midgut content is found to be almost free from fungi [61]. Blocking one of these proteins in termites accelerated and increased fungal infection [58]. β -1,3-glucanase can also hydrolyze callose (β -1,3-glucan). Callose exists in the cell walls of higher plants and plays important roles during plant development. Callose deposition also is induced by a variety of biotic and abiotic stresses such as wounding, pathogen infection, aluminum, abscisic acid, and raised or lowered temperatures [64]. Therefore, it is plausible that *D. v. virgifera* larvae use this enzyme (e.g., encoded by GH16-1, which has significantly high larval expression) to digest callose.

GH5 Family

A short sequence similar to part of GH5 family genes was identified in the *D. v. virgifera* transcriptome (317 bp, 105 aa) (Figure S4A). Among the 51 GH5 subfamilies [65], coleopteran GH5 genes known so far belong to three subfamilies (2, 8, and 10) (Figure S4B). The subfamily 8 gene found in the coffee berry borer *Hypothenemus hampei* (Curculionoidea), however, was shown to be bacterial origin [66]. The short *D. v. virgifera* GH5 sequence is phylogenetically closer to fungal GH5 sequences belonging to the subfamily 12 (Figure S4B). We should, however, note that we

failed to confirm the corresponding sequence in the draft *D. v. virgifera* genome. Furthermore, the expression of this sequence was not confirmed with confidence (Table S4). Therefore, we consider the existence of a GH5 gene in *D. v. virgifera* to be inconclusive.

Absence of GH9 Family in *D. v. virgifera*

GH9 family genes have been identified in insect orders Orthoptera, Blattaria, Phthiraptera, Hemiptera, Coleoptera, and Hymenoptera [11,67]. However, no GH9 candidate sequence was identified in the *D. v. virgifera* transcriptome (Figure 1). GH9 family genes appear to be absent among chrysomelids and curculionids [11,67]. Among beetle species, a GH9 family gene is present in *T. castaneum* (Tenebrionoidea) [25]. We also found a GH9 family gene sequence from the transcriptome of the salt marsh beetle *Pogonus chalceus* (Carabidae, Adephaga). Because *P. chalceus* is placed as the most basal species in Coleoptera [68] (Figure 1), GH9 family genes were likely maintained in the common ancestor of Coleoptera and the lineage leading to the superfamily Tenebrionoidea. GH9 family genes must have been subsequently lost in the common ancestor of Chrysomeloidea and Curculionoidea.

We confirmed that three GH families (GH45, GH48, and GH28) are absent from the transcriptomes of *P. chalceus* [69] and the genome of *T. castaneum* [70] (Figure 1). The loss of GH9 and gain of GH45, GH48, and GH28 families, therefore, can be traced back at least to the common ancestor of chrysomelids and curculionids. Although GH9 and three enzymes (GH48, GH45, and GH28) do not share sequence similarities and have different 3D structural features (CAZy classifies GH48 in GH-M and GH28 in GH-N clans; GH9 and GH45 are not classified), Watanabe and Tokuda [11] suggested, for example, a possible convergent evolution in terms of enzymatic function based on the same substrate specificities (e.g., β -1,4 linkages) with GH9 and GH45 enzymes. GH9 and GH28 enzymes utilize the inverting glycosidase mechanism, which only allows polysaccharide hydrolysis [71]. Thus, their functional similarities may have allowed the laterally acquired genes to replace the role of the lost GH9 enzymes.

GH1 Family

The hydrolysis of cellulose is completed by β -glucosidases, which hydrolyze cellobiose and oligosaccharides to glucose. The GH1 family, the largest of GH families that encode β -glucosidase activities, has been identified widely in insects (e.g., [72–74]). From the *D. v. virgifera* transcriptome, we identified twenty-eight GH1 gene candidates (23 intact and 5 partial; Figure S5A). Multiple copies of GH1 gene candidates were found also in other coleopteran species (e.g., 19 in *D. ponderosae* and 15 in *T. castaneum*) (Figure 1). While multiple GH1 family genes are regularly found in eukaryotic organisms, this family is particularly expanded in coleopteran species. Phylogenetic analysis shows that all insect GH1 family proteins are monophyletic (although bootstrap support is marginal; 76% only by the maximum-likelihood phylogeny) (Figure S5B). Consistent with previous studies, GH1 family proteins form clusters according to each domain of life

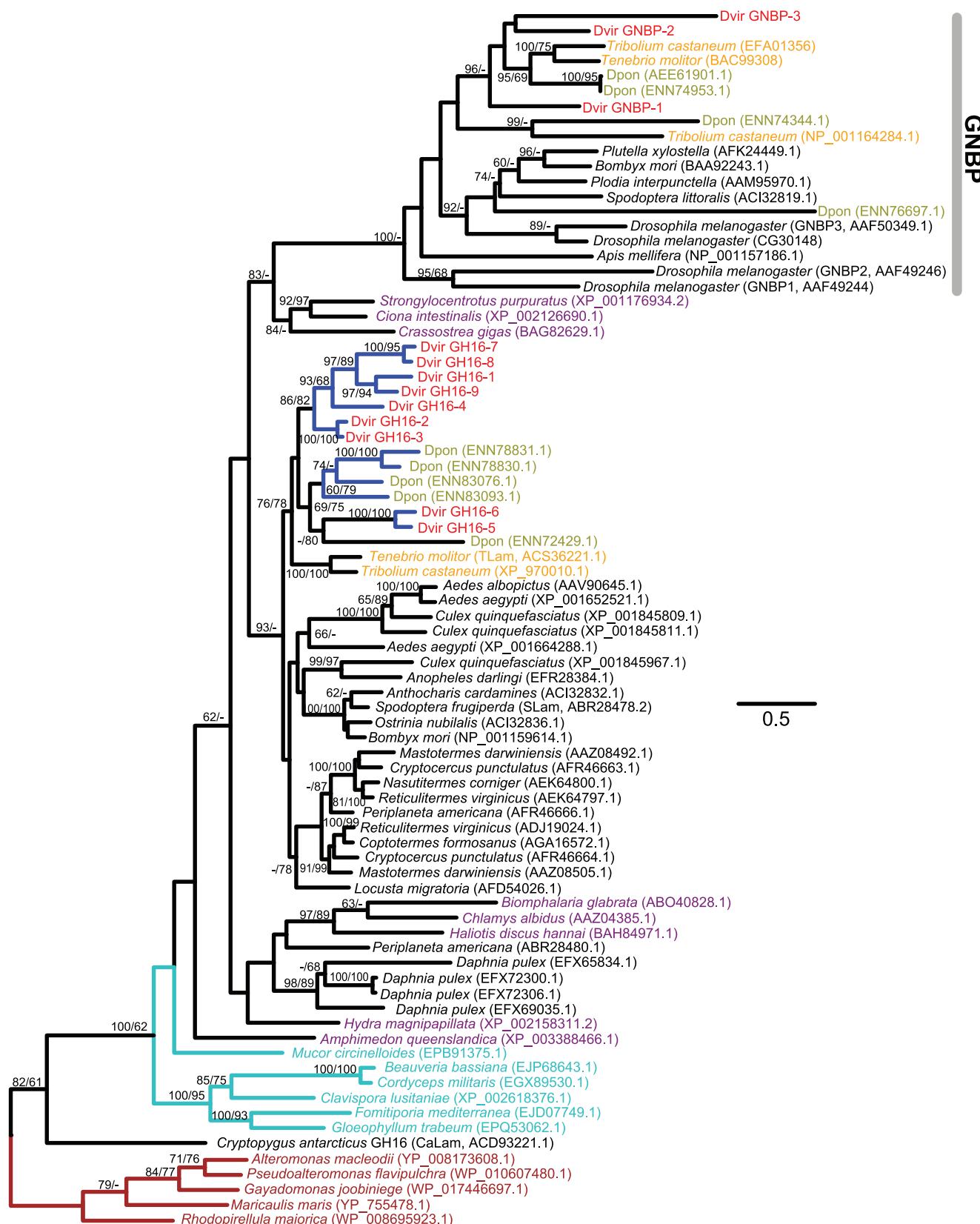


Figure 5. The maximum-likelihood phylogeny of GH16 proteins. Sixteen GH16 protein sequences from four coleopteran species are included. Labels for the coleopteran species belonging to the superfamily Curculionoidea, *D. v. virgifera*, and other beetle sequences are shown in olive, red, and orange, respectively. Their species abbreviations are found in Table S5. Arthropod, other metazoan, fungal (6 chosen from 222 sequences), and bacterial (5 chosen from 977 sequences) sequences are indicated by black, purple, cyan, and brown, respectively. Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining

phylogenies in this order. Supporting values are shown only when higher than 60%. Blue-colored branches indicate the species-specific gene duplications (based on currently available sequences) within a cluster supported by higher than 70% of bootstrap values. The scale bar represents the number of amino acid substitutions per site.

doi:10.1371/journal.pone.0094052.g005

[75,76]. GH1 family expansion appears to have happened independently in many different lineages. For example, four β -glucosidases have been identified in the midgut of the yellow mealworm *T. molitor*, and their enzyme specificities and efficiencies differ slightly [72,77]. The majority of GH1 gene candidates we identified from *D. v. virgifera* have high expression in larvae, especially in the third-instar larval midgut, but not in eggs (Table S4). Therefore, these GH1 gene products likely participate in digesting various plant materials.

GH31 and GH27 Families

In addition to cellulases and other plant cell wall digesting enzymes, we also searched genes encoding GH31 (α -glucosidase) and GH27 (α -galactosidase) families. These enzymes share a common $(\beta/\alpha)_8$ (TIM) barrel catalytic domain and belong to the same GH-D clan [78]. In mosquitoes, an α -glucosidase has been shown to function as a receptor of Bin toxin from *Bacillus sphaericus* as well as Cry11Ba toxin from *B. thuringiensis* subsp. *jegathesan* [79,80]. Ten GH31 and two GH27 family gene candidates were identified from the *D. v. virgifera* transcriptome (Figures S6A and S7A). All of these genes are highly expressed especially in the third-instar larval midgut (Table S4). Both families are found in a wide range of organisms from bacteria to eukaryotes (Figures S6B and S7B). Wheeler *et al.* [81] showed two lepidopteran GH31-related sequences as HGT origin from bacteria (BGIBMGA013995 and Px016165 in Figure S6B). We found no such evidence in search of GH31 family genes in *D. v. virgifera* as well as in *D. ponderosae* transcriptomes.

GH Family Gene Expression

We compared the expression levels of GH family gene candidates we identified from the *D. v. virgifera* transcriptomes between egg and larval samples. Almost all were expressed significantly more in larval stages. We found that the majority of GH45, GH28, GH1, GH31, and GH27 family genes are expressed more in the third-instar larval midgut samples compared to egg and neonate samples, with some genes particularly standing out (GH45-4, GH45-7, GH45-10, GH28-6, GH1-18, GH27-1, GH31-7) (Table S4). Gene expression and enzyme activity of polygalacturonase have been reported from the gut of another corn rootworm species, *Diabrotica undecimpunctata howardi* (spotted cucumber beetle) [82]. GH28 and GH45 family genes are expressed more in the guts of *P. cochleariae* larvae and adults [83]. Polygalacturonases are known to loosen the primary cell wall and make cellulose-hemicellulose network more accessible to enzymatic digestion [84]. With its high number of GH45, GH28, and GH1 genes and their high expression in larval midgut tissue, *D. v. virgifera* may utilize β -1,4-endoglucanase, polygalacturonase, as well as β -glucosidase activities in larval midgut to assist in the digestion of corn root cell walls and in releasing dietary monosaccharides such as glucose.

Horizontal Gene Transfer of GH Family Genes

Our current study indicates that the three GH gene families (GH45, GH48, and GH28) are unique to the two coleopteran superfamilies (Chrysomeloidea and Curculionoidea) and generally absent from other insects except in plant bugs (GH28) and in a springtail (GH45). These results imply that these genes are likely not vertically inherited from the ancestral species but acquired by

HGT events from bacteria or fungi to the common ancestor of chrysomelid and curculionid beetles.

As mentioned before, the GH5 family gene (*HhMANI*) identified from the coffee berry borer *H. hampei* is thought to be bacterial origin [66]. This gene was not found in two other related species, *H. obscurus* (topical nut borer) and *Araecerus fasciculatus* (coffee bean weevil, Anthribidae, Coleoptera). *H. obscurus* is a pest of macadamia nuts but not coffee [85]. *A. fasciculatus* is polyphagous, a common pest of stored food products including coffee [86]. In contrast, *H. hampei* is mainly a coffee pest, although it may not be strictly monophagous [87,88]. Therefore, acquisition of *HhMANI* from bacteria may have made a rapid adaptation possible for *H. hampei* by enabling hydrolysis of galactomannan, the major nutrient source for this species [66]. Other examples of possible HGTs of GH family genes include: GH5 and GH11 family genes in rumen fungi from rumen bacteria *Fibrobacter succinogenes* [43], a GH16 family gene in *C. antarcticus* from bacteria [52], lepidopteran GH31-like genes from *Enterococcus* bacteria [81], and GH11 family genes in *P. cochleariae* from γ -proteobacteria [24]. We also found evidence of several independent HGT events such as fungal GH48 family genes and plant bug GH28 family genes. Although HGT events are often detected in prokaryotes [89], GH families seem to be characterized by frequent HGT events in various animals especially in insects. Such acquisitions followed by frequent duplications of these GH genes must have contributed to these organisms' ability to adapt to novel niches.

Conclusion

We have identified eight GH family genes from the transcriptomes of *D. v. virgifera*. Three GH families (GH45, GH48, and GH28) were likely to have been obtained by HGT events before the divergence of chrysomelid and curculionid beetles. Rapid birth-and-death processes have been also observed among these coleopteran GH family genes. A large number of GH family enzymes owing to their lineage-specific duplications in *D. v. virgifera* could have contributed to the successful adaptation to its niche by providing more efficient hydrolyzation of corn cell walls.

Materials and Methods

Sample Collection and Preparation

Eggs. Two thousands freshly hatched non-diapause *D. v. virgifera* eggs (ten Petri dishes, ~200 eggs/Petri dish) were purchased from Crop Characteristics, Inc. (Farmington, Minnesota, USA). All ten Petri dishes were wrapped with aluminum foil and placed in a growth chamber for incubation at 27°C. One Petri dish was removed from incubator on each day and the eggs were isolated with a 60-mesh sieve. Briefly, the soil with eggs were rinsed with tap water until soil was removed completely, and the isolated eggs were washed with double distilled water before being transferred into a 1.7 ml centrifuge tube. The water was removed with pipette and eggs were snap-frozen in liquid nitrogen and stored in -80°C freezer. All other Petri dishes were processed in the same way until the day 10.

Neonates. A Petri dish containing 10,000 eggs was purchased from Crop Characteristics, Inc. (Farmington, Minnesota, USA) and placed in a growth chamber at 27°C with LD 16:8 photoperiod until hatching. The eggs were isolated from soil with methods described above. The clean eggs were rinsed with double

distilled water three times before transferring to a new egg Petri dish (60×15 mm) with moistened filter paper (42.5 mm). Finally, the Petri dish was placed back to the same growth chamber for more neonates to hatch. The freshly hatched (less than 24 hrs old) neonates were collected, snap-frozen in liquid nitrogen, and stored at -80°C freezer.

Preparation of midgut from third instar larvae. Fifty third-instar larvae purchased from Crop Characteristics, Inc. (Farmington, Minnesota, USA) were dissected for midgut tissue under dissection microscope. Briefly, the head, thorax, and last two segments of abdomen were removed with a scalpel and the midgut was pulled from the carcass with forceps. The fat body and other liquids were carefully removed by pulling the midgut on a clean Kimtech Science Precision Wipes Tissue Wipers (Fisher). The midgut was then opened longitudinally and gut contents were removed by rinsing 3 times with 1X PBS buffer (pH 7.4). The clean midgut tissue was snap-frozen in liquid nitrogen and saved at -80°C until RNA extraction. Sixteen midgut tissues were pooled as one replicate with 3 replicates in total.

RNA extraction. Approximately 35 mg of pooled midgut tissue was used directly in a single RNA preparation while ~32 pooled neonates were used for single RNA extraction. For egg RNA preparation, ~35 mg of pooled eggs from day 1 to day 10 were used in single RNA preparation. Three replicates for eggs and midgut dissection and six replicates for neonates were prepared. The total RNAs were extracted with RNeasy mini kit (Qiagen, Cat. 74104) and treated with RNase-free DNase set (Qiagen, Cat. 79254) for potential genomic DNA contamination by following the manufacturers' instructions. The quality and quantity of RNA were evaluated on 1% agarose gel and NanoDrop 1000 (Thermo) for further analysis.

Next Generation Sequencing and Assembly of *D. v. virgifera* Transcriptomes

Next generation sequencing. The 454 pyrosequencing experiments of larval midgut samples were completed using Roche GS-FLX titanium sequencer at the Core for Applied Genomics and Ecology, University of Nebraska-Lincoln. The transcriptome sequencing for the egg and larval midgut samples with an insert size of 300 bp was done on Illumina Genome Analyzer II platform at the Center for Biotechnology, University of Nebraska-Lincoln. The neonate samples were sequenced with an insert size of 500 bp on Illumina HiSeq2000 system at the Durham Research Center, University of Nebraska Medical Center. In total, 16.6 gigabases (Gb) (read length 75 bp) of egg RNA, 33 Gb (read length 75 bp) of larval midgut RNA, and 662 Gb (read length 101 bp) of neonate RNA were sequenced. All 454 and Illumina reads were deposited into the NCBI Sequence Read Archive (SRA) under the accession number SRR037561.

Filtering of low quality reads. Because sequencing errors can cause difficulties for the assembly algorithm, we applied a stringent quality filter process. For 454 reads, the adapter and poly(A/T) sequences were trimmed using PRINSEQ [90]. 454 reads that have abnormal read length (<50 bp or >1000 bp) or where the average quality was less than 20 were removed. The Illumina paired-end reads that did not have the minimum quality score (20 per base for egg and midgut samples or 30 per base for neonate samples) across the whole read were removed using PRINSEQ [90] and Sickle (ver. 1.2) [91]. Note that the quality scores of 20 (Q20) and 30 (Q30) correspond to 1% and 0.1% expected error rates, respectively. We removed all Illumina reads that have unknown nucleotide 'N'.

de novo transcriptome assembly. After the filtering process, we performed *de novo* transcriptome assembly for each

of three samples. We used four different short read assemblers: Newbler (ver. 2.5) (Roche, 454 Life Sciences; used only for 454 read assembly), Mira (ver. 3.4.0) [92], Velvet/Oasis (ver. 1.2.03) [93], and Trinity (release 2013-02-25) [28]. The k-mer size of 25 was used for all programs. Mira could be used only for 454 read assembly from the third instar larval samples and for the Illumina read assembly from the egg samples due to the large memory requirement. The results of these assemblies are summarized in Table S1. The number of assembled transcripts varied among the different assemblers, ranging from 37,181 by Trinity to 165,361 for Velvet/Oasis for 454 reads (larval midgut sample) and from 56,135 by Velvet/Oasis to 72,638 by Trinity for Illumina reads (egg sample). The average length and N50 of contigs were generally longer with the Trinity assembly (Tables S1 and S2). Results of NCBI BLAST similarity search (blastx, ver. 2.2.26+) [94,95] against the UniProt protein database (<http://www.uniprot.org>) [96] showed that fractions of contigs that had highly significant hits (E-value $\leq 10^{-100}$) were larger with the Trinity (18.9%) and Velvet/Oasis assemblies (~19.4%) than the Mira assembly (11%) although the difference was not significant ($P>0.5$ by t-test between Trinity and Mira) (Figure S8). Note that Zhao *et al.* [97] showed the highest accuracy with Trinity among methods specialized in *de novo* transcriptome assemblies such as SOAPdenovo [98], ABySS [99], Velvet/Oasis, and Trinity (they did not include Mira in their comparison). We also attempted the hybrid assemblies using two different sequencing platforms (454 and Illumina Genome Analyzer II) for the third instar larval midgut sample as well as for the pooled egg and third instar larval midgut samples (Table S3). Furthermore, we performed assembly using the dataset pooled from egg (produced by Illumina Genome Analyzer II), third instar larval midgut (produced by Illumina Genome Analyzer II), and neonate samples (produced by Illumina HiSeq2000). Among all of these assemblies, the Trinity assembly using the pooled Illumina dataset had the longest average length of contigs and N50, even longer than the hybrid assemblies including 454 reads (Table 1). With this assembly, more GH family gene candidates were also identified. Therefore, we used this Trinity assembly using the pooled dataset as the most inclusive "combined *D. v. virgifera* transcriptome" for all our further studies.

Sequence Analysis

Gene expression analysis. To compare the gene expression levels, the paired-end reads were mapped onto our combined *D. v. virgifera* transcriptome using bowtie (ver. 1.0.0) [100] with 0 mismatch. The numerical values of gene expression were measured by RPKM (reads per kilobase per million mapped reads) to normalize for the number of sequencing reads and total read length [101]. RPKM values above 0.3 [102] as well as having more than 10 reads was used as the threshold for gene expression.

Identification of GH family genes from the *D. v. virgifera* transcriptome. Previously reported insect, especially coleopteran, GH family gene sequences were obtained for GH45 [22,23,31], for GH48 [22,39], for GH28 [22], for GH9 [25], for GH5 [66], for GH11 [24], for GH16 [61], for GH1 [72,73], for GH27 (XP_973339.2; this entry is incorrectly shown to be a GH31 family enzyme), and for GH31 [81] (see Table S5). Using these sequences as initial queries, we searched GH family gene candidates against our combined *D. v. virgifera* transcriptome using NCBI BLAST (tblastn, ver. 2.2.26+) [94,95]. The initial E-value threshold (1×10^{-6}) used was rather lenient and chosen to be inclusive of all true positives even though some false positives from non-target genes may have been included. With recurrent phylogenetic analyses and BLAST (blastp) similarity searches, we identified protein sequences for each GH family. Reciprocal

similarity search (using blastp) was further performed against the NCBI NR protein database to confirm GH family associations. All *D. v. virgifera* GH gene candidate sequences were also confirmed by BLAST (tblastn, E-value $\leq 10^{-30}$) similarity search against the draft *D. v. virgifera* genome sequence (Hugh M. Robertson, personal communication). The criterion we used to identify alternative spliced isoforms was to have a more than 60 bp of 100% identical region among the candidate sequences. We did not find any potential alternative spliced isoforms for the GH gene sequences we identified.

GH family gene search. In order to identify GH family sequences from a wide range of insects and other organisms, we performed BLAST similarity searches (blastp, E-value $\leq 10^{-30}$) using *D. v. virgifera* and other representative GH gene sequences as queries against the NCBI NR protein database (<http://www.ncbi.nlm.nih.gov>) as well as against ten insect genomes (*Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, *Bombyx mori*, *Apis mellifera*, *Nasonia vitripennis*, *Solenopsis invicta*, *Ixodes scapularis*, *Rhodnius prolixus*, and *Acyrthosiphon pisum*). For all BLAST similarity searches, in order to obtain comparable E-values, the database size was set to 1.1×10^{10} (using the ‘-dbsize’ option), which is based on the database size equivalent to the NCBI NR database.

Multiple sequence alignments and phylogenetic analysis. Multiple alignments of GH family protein sequences were generated using MAFFT (ver. 7.050b) with the L-INS-i algorithm, which uses a consistency-based objective function and local pairwise alignment with affine gap costs [103]. Phylogenetic relationships were reconstructed by the maximum-likelihood method using RAxML (ver. 7.0.4) [104] with the PROT GAM-MAJTT substitution model (JTT matrix with gamma-distributed rate variation). The neighbor-joining phylogenies [105] were reconstructed by using neighbor of the Phylip package (ver. 3.69) [106]. The protein distances were estimated using protdist of the Phylip package with the JTT model. Non-parametric bootstrapping with 1000 pseudoreplicates [107] was used to estimate the confidence of branching patterns. FigTree (<http://tree.bio.ed.ac.uk/software/figtree>) was used to display the phylogenetic trees.

The nucleotide sequences of all GH family genes identified and the alignments used for the phylogenetic analysis in this study are available from our website: <http://bioinfolab.unl.edu/emlab/GH/>.

Supporting Information

Figure S1 GH family gene sequences identified from the *D. v. virgifera* transcriptome. Amino acid sequences of GH45 (A), GH48 (B), and GH28 (C) are shown in alignments. The labels for partial sequences are shown in italics. Potential residues for the catalytic nucleophile and the proton donor are highlighted with magenta and green, respectively (based on Sakamoto and Toyohara, 2009, *Comp Biochem Physiol B* **152**: 390; Parsiegla *et al.*, 2008, *J Mol Biol* **375**: 499; van Santen *et al.*, 1999, *J Biol Chem* **274**: 30474).

(PDF)

Figure S2 The maximum-likelihood phylogeny of GH45 family proteins. Labels for the coleopteran species belonging to the superfamily Curculionoidea are olive-colored and all other coleopteran sequences colored in black belong to the superfamily Chrysomeloidea. *D. v. virgifera* sequences are shown in red. Their species name abbreviations are found in Table S5. Other sequences are colored as follows: mollusks (purple), *Cryptopygus antarcticus* (Collembola, black), *Hypsibius dujardini* (Tardigrada, black), protists (dark green), plant-parasitic nematodes (grey),

fungi (cyan), and bacteria (brown). The scale bar represents the number of amino acid substitutions per site.

(PDF)

Figure S3 Multiple alignments of *D. v. virgifera* GH16 family protein sequences. A. GH16 family proteins identified from the *D. v. virgifera* transcriptome. B. The active site region sequences. GNBP sequences are boxed. The catalytic nucleophile and proton donor residues are highlighted with magenta and green, respectively (based on Viladot *et al.* 1998, *Biochemistry* **34**: 11332). C. The N-terminal conserved domain sequences of GNBPs. Residue shown to be within hydrogen-binding distances and involved in hydrophilic interaction with lamitrihexoses from *Plodia interpunctella* and *Bombyx mori* proteins are highlighted with yellow and Arg's involved in binding of triplex β-glucan are highlighted in light blue (based on Kanagawa *et al.*, 2011, *J Biol Chem* **286**: 19158).

(PDF)

Figure S4 Multiple alignment of the potential GH5 family protein sequence identified from *D. v. virgifera* with four fungal GH5 proteins (A) and the maximum-likelihood phylogeny including other known GH5 family proteins (B). The potential amino acid residues for the catalytic nucleophile and catalytic proton donor are highlighted with magenta and green in the alignment, respectively (based on Larsson *et al.* 2006, *J Mol Biol* **357**: 1500). Coleopteran proteins included in the phylogeny are found in Table S5. The *D. v. virgifera* sequence is shown in red. The GH5 protein sequences are classified into subfamilies according to Aspeborg *et al.* (2012, *BMC Evol Biol* **12**: 186). Bacterial, plant, fungal, and nematode sequences are indicated by brown, green, cyan, and grey. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Only bootstrap values higher than 70% are shown.

(PDF)

Figure S5 GH1 family gene sequences identified from the *D. v. virgifera* transcriptome (A) and the maximum-likelihood phylogeny of GH1 family proteins (B). In the alignment, the labels for partial sequences are shown in italics. Potential residues for the catalytic nucleophile and the proton are highlighted with magenta and green, respectively (based on Marana *et al.*, 2001, *Biochim Biophys Acta* **1545**: 41; Scharf *et al.* 2010, *Insect Biochem Mol Biol* **40**: 611). In the phylogeny, labels for the coleopteran species belonging to the superfamily Curculionoidea, *D. v. virgifera*, and other beetle sequences are shown in olive, red, and orange, respectively. Their species abbreviations are found in Table S5. Arthropod, other metazoan, nematode, fungal, plant, and bacterial sequences are indicated by black, purple, grey, cyan, green, and brown, respectively. The numbers at internal branches show the bootstrap support values (%) for the maximum likelihood and neighbor-joining phylogenies in this order. Supporting values are shown only when higher than 60%. The scale bar represents the number of amino acid substitutions per site.

(PDF)

Figure S6 GH31 family gene sequences identified from the *D. v. virgifera* transcriptome (A) and the maximum-likelihood phylogeny of GH31 family proteins (B). In the alignment, the labels for partial sequences are shown in italics. In the phylogeny, labels for the coleopteran species belonging to the superfamily Curculionoidea, *D. v. virgifera*, and other beetle sequences are shown in olive, red, and orange, respectively. Their species abbreviations are found in Table S5. Arthropod, other

metazoan, nematode, fungal, plant, and bacterial sequences are indicated by black, purple, grey, cyan, green, and brown, respectively. The accession numbers shown in parenthesis are from NCBI except for: BGIBMGA012077-PA and BGIBMGA013995 from SilkDB (<http://www.silkdb.org>), DPOGS202361 from MonarchBase (<http://monarchbase.umassmed.edu>), and Px016165 from Diamondback moth Genome Database (<http://59.79.254.1/DBM/>). The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Supporting values are shown only when higher than 60%. The scale bar represents the number of amino acid substitutions per site.

(PDF)

Figure S7 GH27 family gene sequences identified from the *D. v. virgifera* transcriptome (A) and the maximum-likelihood phylogeny including representative GH27 family proteins (B). Labels for the coleopteran species belonging to the superfamily Curculionoidea, *D. v. virgifera*, and other beetle sequences are shown in olive, red, and orange, respectively. Their species abbreviations are found in Table S5. Arthropod, other metazoan, nematode, fungal, plant, and bacterial sequences are indicated by black, purple, grey, cyan, green, and brown, respectively. Bacterial sequences were used as outgroups. The numbers at internal branches show the bootstrap support values (%) for the maximum-likelihood and neighbor-joining phylogenies in this order. Supporting values are shown only when higher than 60%. The scale bar represents the number of amino acid substitutions per site.

(PDF)

Figure S8 The distribution of E-values obtained from blastx similarity search against the UniProt protein database using the assemblies generated by three programs using the *D. v. virgifera* egg samples. The numbers of contigs are 18,173 in Mira (blue), 11,035 in Trinity

References

- Wei H, Xu Q, Taylor LE 2nd, Baker JO, Tucker MP, et al. (2009) Natural paradigms of plant cell wall degradation. *Curr Opin Biotechnol* 20: 330–338.
- Tomme P, Warren RAJ, Gilkes NR (1995) Cellulose hydrolysis by bacteria and fungi. In: Poole RK, editor. *Advances in Microbial Physiology*: Academic Press. 1–81.
- Kamm B, Kamm M (2004) Principles of biorefineries. *Appl Microbiol Biotechnol* 64: 137–145.
- Wyman CE (2003) Potential synergies and challenges in refining cellulosic biomass to fuels, chemicals, and power. *Biotechnol Prog* 19: 254–262.
- Zhang Y-HP, Lynd LR (2005) Cellulose utilization by *Clostridium thermocellum*: bioenergetics and hydrolysis product assimilation. *Proc Natl Acad Sci USA* 102: 7321–7325.
- Cleveland LR (1924) The physiological and symbiotic relationships between the intestinal protozoa of termites and their host, with special reference to *Reticulitermes flavipes* Kollar. *Biol Bull Mar Biol Lab* 46: 117–227.
- Martin M (1991) The evolution of cellulose digestion in insects. *Philos Trans R Soc Lond Ser B* 333: 281–288.
- Breznak JA, Brune A (1994) Role of microorganisms in the digestion of lignocellulose by termites. *Annu Rev Entomol* 39: 453–487.
- Watanabe H, Noda H, Tokuda G, Lo N (1998) A cellulase gene of termite origin. *Nature* 394: 330–331.
- Smant G, Stokkermans JP, Yan Y, de Boer JM, Baum TJ, et al. (1998) Endogenous cellulases in animals: isolation of beta-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc Natl Acad Sci USA* 95: 4906–4911.
- Watanabe H, Tokuda G (2010) Cellulolytic systems in insects. *Annu Rev Entomol* 55: 609–632.
- Calderón-Cortés N, Quesada M, Watanabe H, Cano-Camacho H, Oyama K (2012) Endogenous plant cell wall digestion: a key mechanism in insect evolution. *Annu Rev Ecol Evol Syst* 43: 45–71.
- Nakashima K, Watanabe H, Saitoh H, Tokuda G, Azuma JI (2002) Dual cellulose-digesting system of the wood-feeding termite, *Coptotermes formosanus* Shiraki. *Insect Biochem Mol Biol* 32: 777–784.
- Zhou X, Smith JA, Oi FM, Koehler PG, Bennett GW, et al. (2007) Correlation of cellulase gene expression and celullolytic activity throughout the gut of the termite *Reticulitermes flavipes*. *Gene* 395: 29–39.
- Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450: 560–565.
- Tokuda G, Watanabe H (2007) Hidden cellulases in termites: revision of an old hypothesis. *Biol Lett* 3: 336–339.
- Tokuda G, Watanabe H, Lo N (2007) Does correlation of cellulase gene expression and celullolytic activity in the gut of termite suggest synergistic collaboration of cellulases? *Gene* 401: 131–134.
- Scharf ME, Karl ZJ, Sethi A, Boucias DG (2011) Multiple levels of synergistic collaboration in termite lignocellulose digestion. *PloS one* 6: e21709.
- Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* 37: D233–D238.
- Ketudat Cairns JR, Esen A (2010) β -Glucosidases. *Cell Mol Life Sci* 67: 3389–3405.
- Chang CJ, Wu CP, Lu SC, Chao AL, Ho TH, et al. (2012) A novel exocellulase from white-spotted longhorn beetle (*Anoplophora malasiaca*). *Insect Biochem Mol Biol* 42: 629–636.
- Pauchet Y, Wilkinson P, Chauhan R, ffrench-Constant RH (2010) Diversity of beetle genes encoding novel plant cell wall degrading enzymes. *PloS one* 5: e15635.
- Calderón-Cortés N, Watanabe H, Cano-Camacho H, Zavala-Páramo G, Quesada M (2010) cDNA cloning, homology modelling and evolutionary insights into novel endogenous cellulases of the borer beetle *Oncideres abomarginata chamele* (Cerambycidae). *Insect Mol Biol* 19: 323–336.
- Pauchet Y, Heckel DG (2013) The genome of the mustard leaf beetle encodes two active xylenases originally acquired from bacteria through horizontal gene transfer. *Proc R Soc B* 280: 20131021.

(red), and 9843 in Velvet/Oasis (green). E-values are shown as $-\log_{10}$ (E-value) except for E-value = 0. Note that there is no significant difference between Trinity and Mira (*t*-test $P > 0.5$ for both $E\text{-value} \leq 10^{-100}$ and for all E-values).

(PDF)

Table S1 Summary statistics for *D. v. virgifera* transcriptome sequencing and assembly.

(PDF)

Table S2 Summary of *D. v. virgifera* transcriptome sequencing and assemblies.

(PDF)

Table S3 Summary statistics for hybrid and pooled-data assembly of *D. v. virgifera* transcriptome.

(PDF)

Table S4 Expression analysis of the *D. v. virgifera* GH family genes identified in this study.

(PDF)

Table S5 Coleopteran GH family gene sequences used in this study.

(PDF)

Acknowledgments

We would like to thank Drs. Kim Walden and Hugh M. Robertson (University of Illinois at Urbana-Champaign, USA) for letting us use the draft genome sequence of *D. v. virgifera* and Steven M. Van Belleghem for the transcriptome of *P. chalceus*.

Author Contributions

Conceived and designed the experiments: BDS SE EM AKB. Performed the experiments: SE YP HW. Analyzed the data: SE EM HW. Contributed reagents/materials/analysis tools: YP RF AV. Wrote the paper: SE EM BDS. Prepared library for 454 sequencing: YP RF.

25. Willis JD, Oppert B, Oppert C, Klingeman WE, Jurat-Fuentes JL (2011) Identification, cloning, and expression of a GHF9 cellulase from *Tribolium castaneum* (Coleoptera: Tenebrionidae). *J Insect Physiol* 57: 300–306.

26. Valencia A, Alves AP, Siegfried BD (2013) Molecular cloning and functional characterization of an endogenous endoglucanase belonging to GHF45 from the western corn rootworm, *Diabrotica virgifera virgifera*. *Gene* 513: 260–267.

27. Siegfried BD, Waterfield N, Ffrench-Constant RH (2005) Expressed sequence tags from *Diabrotica virgifera virgifera* midgut identify a coleopteran cadherin and a diversity of cathepsins. *Insect Mol Biol* 14: 137–143.

28. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* 29: 644–652.

29. Girard C, Jouanin L (1999) Molecular cloning of a gut-specific chitinase cDNA from the beetle *Phaenotrichus cochleariae*. *Insect Biochem Mol Biol* 29: 549–556.

30. Eigenheer AL, Keeling CI, Young S, Tittiger C (2003) Comparison of gene representation in midguts from two phytophagous insects, *Bombyx mori* and *Ips pini*, using expressed sequence tags. *Gene* 316: 127–136.

31. Lee SJ, Kim SR, Yoon HJ, Kim I, Lee KS, et al. (2004) cDNA cloning, expression, and enzymatic activity of a cellulase from the mulberry longicorn beetle, *Apriona germari*. *Comp Biochem Physiol B: Biochem Mol Biol* 139: 107–116.

32. Gao Y, Yu Y, Luan Y-X (2008) Phylogenetic relationships of basal hexapods reconstructed from nearly complete 18S and 28S rRNA gene sequences. *Zool Sci* 25: 1139–1145.

33. Davison A, Blaxter M (2005) Ancient origin of glycosyl hydrolase family 9 cellulase genes. *Mol Biol Evol* 22: 1273–1284.

34. Ohitoko K, Ohkuma M, Moriya S, Inoue T, Usami R, et al. (2000) Diverse genes of cellulase homologues of glycosyl hydrolase family 45 from the symbiotic protists in the hindgut of the termite *Reticulitermes speratus*. *Extremophiles* 4: 343–349.

35. Todaka N, Inoue T, Saita K, Ohkuma M, Nalepa CA, et al. (2010) Phylogenetic analysis of cellulolytic enzyme genes from representative lineages of termites and a related cockroach. *PLoS One* 5: e8636.

36. Li L, Frohlich J, Pfeiffer P, Konig H (2003) Termite gut symbiotic archaezoa are becoming living metabolic fossils. *Eukaryot Cell* 2: 1091–1098.

37. Xu B, Janson J-C, Sellos D (2001) Cloning and sequencing of a molluscan endo- β -1,4-glucanase gene from the blue mussel, *Mytilus edulis*. *Eur J Biochem* 268: 3718–3727.

38. Harada Y, Hosoiri Y, Kuroda R (2004) Isolation and evaluation of dextral-specific and dextral-enriched cDNA clones as candidates for the handedness-determining gene in a freshwater gastropod, *Lymnaea stagnalis*. *Dev Genes Evol* 214: 159–169.

39. Fujita K, Shimomura K, Yamamoto K, Yamashita T, Suzuki K (2006) A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. *Biochem Biophys Res Commun* 345: 502–507.

40. Keeling CI, Henderson H, Li M, Yuen M, Clark EL, et al. (2012) Transcriptome and full-length cDNA resources for the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major insect pest of pine forests. *Insect Biochem Mol Biol* 42: 525–536.

41. Berger E, Zhang D, Zverlov VV, Schwarz WH (2007) Two noncellulosomal cellulases of *Clostridium thermocellum*, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically. *FEMS Microbiol Lett* 268: 194–201.

42. Ljungdahl LG (2008) The cellulase/hemicellulase system of the anaerobic fungus orpinomyces PC-2 and aspects of its applied use. *Ann N Y Acad Sci* 1125: 308–321.

43. Garcia-Vallvé S, Romeu A, Palau J (2000) Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol Biol Evol* 17: 352–361.

44. Hung YL, Chen HJ, Liu JC, Chen YC (2012) Catalytic efficiency diversification of duplicate beta-1,3–1,4-glucanases from *Neocallimastix patriciarum* J11. *Appl Environ Microbiol* 78: 4294–4300.

45. Keeling C, Yuen M, Liao N, Docking T, Chan S, et al. (2013) Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biology* 14: R27.

46. Celorio-Mancera MP, Allen ML, Powell AL, Ahmadi H, Salemi MR, et al. (2008) Polygalacturonase causes lygus-like damage on plants: cloning and identification of western tarnished plant bug (*Lygus hesperus*) polygalacturonases secreted during feeding. *Arthropod-Plant Interactions* 2: 215–225.

47. Allen ML, Mertens JA (2008) Molecular cloning and expression of three polygalacturonase cDNAs from the tarnished plant bug, *Lygus lineolaris*. *J Insect Sci* 8: 1–14.

48. Pauchet Y, Freitak D, Heidel-Fischer HM, Heckel DG, Vogel H (2009) Immunity or digestion: glucanase activity in a glucan-binding protein family from Lepidoptera. *J Biol Chem* 284: 2214–2224.

49. Bragatto I, Genta FA, Ribeiro AF, Terra WR, Ferreira C (2010) Characterization of a beta-1,3-glucanase active in the alkaline midgut of *Spodoptera frugiperda* larvae and its relation to beta-glucan-binding proteins. *Insect Biochem Mol Biol* 40: 861–872.

50. Hughes AL (2012) Evolution of the betaGRP/GNBP/beta-1,3-glucanase family of insects. *Immunogenetics* 64: 549–558.

51. Kovalchuk SN, Bakunina IY, Burtseva YV, Emelyanenko VI, Kim NY, et al. (2009) An endo-(1→3)-beta-D-glucanase from the scallop *Chlamys albida*: catalytic properties, cDNA cloning and secondary-structure characterization. *Carbohydrate Research* 344: 191–197.

52. Song JM, Nam K, Sun YU, Kang MH, Kim CG, et al. (2010) Molecular and biochemical characterizations of a novel arthropod endo-beta-1,3-glucanase from the Antarctic springtail, *Cryptopygus antarcticus*, horizontally acquired from bacteria. *Comp Biochem Physiol B Biochem Mol Biol* 155: 403–412.

53. Kim Y-S, Ryu J-H, Han S-J, Choi K-H, Nam K-B, et al. (2000) Gram-negative bacteria-binding protein, a pattern recognition receptor for lipopolysaccharide and beta-1,3-glucan that mediates the signaling for the induction of innate immune genes in *Drosophila melanogaster* cells. *J Biol Chem* 275: 32721–32727.

54. Pili-Floury S, Leulier F, Takahashi K, Saigo K, Samain E, et al. (2004) In vivo RNA interference analysis reveals an unexpected role for GNBP1 in the defense against Gram-positive bacterial infection in *Drosophila* adults. *J Biol Chem* 279: 12848–12853.

55. Zhang R, Cho HY, Kim HS, Ma YG, Osaki T, et al. (2003) Characterization and properties of a 1,3-beta-D-glucan pattern recognition protein of *Tenebrio molitor* larvae that is specifically degraded by serine protease during prophenoloxidase activation. *J Biol Chem* 278: 42072–42079.

56. Ma C, Kanost MR (2000) A beta-1,3-glucan recognition protein from an insect, *Manduca sexta*, agglutinates microorganisms and activates the phenoloxidase cascade. *J Biol Chem* 275: 7505–7514.

57. Ochiai M, Ashida M (2000) A pattern-recognition protein for beta-1,3-glucan. The binding domain and the cDNA cloning of beta-1,3-glucan recognition protein from the silkworm, *Bombyx mori*. *J Biol Chem* 275: 4995–5002.

58. Bulmer MS, Bacheler I, Raman R, Rosengaus RB, Sasisiekharan R (2009) Targeting an antimicrobial effector function in insect immunity as a pest control strategy. *Proc Natl Acad Sci USA* 106: 12652–12657.

59. Kanagawa M, Satoh T, Ikeda A, Adachi Y, Ohno N, et al. (2011) Structural insights into recognition of triple-helical beta-glucans by an insect fungal receptor. *J Biol Chem* 286: 29158–29165.

60. Mishima Y, Quintin J, Aimananda V, Kellenberger C, Coste F, et al. (2009) The N-terminal domain of *Drosophila* Gram-negative binding protein 3 (GNBP3) defines a novel family of fungal pattern recognition receptors. *J Biol Chem* 284: 28687–28697.

61. Genta FA, Bragatto I, Terra WR, Ferreira C (2009) Purification, characterization and sequencing of the major beta-1,3-glucanase from the midgut of *Tenebrio molitor* larvae. *Insect Biochem Mol Biol* 39: 861–874.

62. Genta FA, Dumont AF, Marana SR, Terra WR, Ferreira C (2007) The interplay of processivity, substrate inhibition and a secondary substrate binding site of an insect exo-beta-1,3-glucanase. *Biochim Biophys Acta* 1774: 1079–1091.

63. Genta FA, Terra WR, Ferreira C (2003) Action pattern, specificity, lytic activities, and physiological role of five digestive beta-glucanases isolated from *Periplaneta americana*. *Insect Biochem Mol Biol* 33: 1085–1097.

64. Chen XY, Kim JY (2009) Callose synthesis in higher plants. *Plant Signal Behav* 4: 489–492.

65. Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol Biol* 12: 186.

66. Acuña R, Padilla BE, Flórez-Ramos CP, Rubio JD, Herrera JC, et al. (2012) Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci USA* 109: 4197–4202.

67. Fischer R, Ostafe R, Twyman RM (2013) Cellulases from insects. Advances in biochemical engineering/biotechnology 136: 51–64.

68. Hunt T, Bergsten J, Levkanicova Z, Papadopoulou A, John OS, et al. (2007) A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science* 318: 1913–1916.

69. Van Belleghem SM, Roelofs D, Van Houdt J, Hendrikx F (2012) *In silico* transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalcus* (Coleoptera, Carabidae). *PLoS one* 7: e42605.

70. Tribolium Genome Sequencing Consortium (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452: 949–955.

71. Sinnott M (1990) Catalytic mechanisms of enzymatic glycosyl transfer. *Chem Rev* 90: 1171–1202.

72. Ferreira AH, Marana SR, Terra WR, Ferreira C (2001) Purification, molecular cloning, and properties of a beta-glycosidase isolated from midgut lumen of *Tenebrio molitor* (Coleoptera) larvae. *Insect Biochem Mol Biol* 31: 1065–1076.

73. Marana SR, Jacobs-Lorena M, Terra WR, Ferreira C (2001) Amino acid residues involved in substrate binding and catalysis in an insect digestive beta-glycosidase. *Biochim Biophys Acta* 1545: 41–52.

74. Scharf ME, Kovaleva ES, Jadhao S, Campbell JH, Buchman GW, et al. (2010) Functional and translational analyses of a beta-glucosidase gene (glycosyl hydrolase family 1) isolated from the gut of the lower termite *Reticulitermes flavipes*. *Insect Biochem Mol Biol* 40: 611–620.

75. Mian IS (1998) Sequence, structural, functional, and phylogenetic analyses of three glycosidase families. *Blood cells, molecules & diseases* 24: 83–100.

76. Marques AR, Coutinho PM, Videira P, Fialho AM, Sa-Correia I (2003) *Sphingomonas paucimobilis* beta-glucosidase Bgl1: a member of a new bacterial subfamily in glycoside hydrolase family 1. *The Biochemical journal* 370: 793–804.

77. Ferreira AH, Terra WR, Ferreira C (2003) Characterization of a beta-glycosidase highly active on disaccharides and of a beta-galactosidase from *Tenebrio molitor* midgut lumen. *Insect Biochem Mol Biol* 33: 253–265.

78. Ernst HA, Lo Leggio L, Willemoes M, Leonard G, Blum P, et al. (2006) Structure of the *Sulfolobus solfataricus* alpha-glucosidase: implications for domain conservation and substrate recognition in GH31. *J Mol Biol* 358: 1106–1124.

79. Opota O, Charles JF, Warot S, Pauron D, Darboux I (2008) Identification and characterization of the receptor for the *Bacillus sphaericus* binary toxin in the malaria vector mosquito, *Anopheles gambiae*. *Comp Biochem Physiol B Biochem Mol Biol* 149: 419–427.

80. Zhang Q, Hua G, Bayyareddy K, Adang MJ (2013) Analyses of alpha-amylase and alpha-glucosidase in the malaria vector mosquito, *Anopheles gambiae*, as receptors of Cry11Ba toxin of *Bacillus thuringiensis* subsp. *jegathesan*. *Insect Biochem Mol Biol* 43: 907–915.

81. Wheeler D, Redding AJ, Werren JH (2013) Characterization of an ancient lepidopteran lateral gene transfer. *PLoS one* 8: e59262.

82. Shen Z, Denton M, Mutti N, Pappan K, Kanost MR, et al. (2003) Polygalacturonase from *Staphylococcus oryzae*: possible horizontal transfer of a pectinase gene from fungi to weevils. *J Insect Sci* 3: 24.

83. Kirsch R, Wiedl N, Vogel H, Svatos A, Heckel D, et al. (2012) Combining proteomics and transcriptome sequencing to identify active plant-cell-wall-degrading enzymes in a leaf beetle. *BMC Genomics* 13: 587.

84. Juge N (2006) Plant protein inhibitors of cell wall degrading enzymes. *Trends Plant Sci* 11: 359–367.

85. Waller JM, Bigger M, Hillocks RJ (2007) Postharvest and processing pests and microbial problems. In: Waller JM, Bigger M, Hillocks RJ, editors. *Coffee Pests, Diseases and Their Management*. CABI, Wallingford, UK. 325–335.

86. Valentine BD (2005) The scientific name of the coffee bean weevil and some additional bibliography (Coleoptera: Anthribidae: *Araecerus Schönherr*). *Insecta Mundi* 19: 247–253.

87. Damon A (2000) A review of the biology and control of the coffee berry borer, *Hypothenemus hampei* (Coleoptera: Scolytidae). *Bull Entomol Res* 90: 453–465.

88. Vega FE, Davis AP, Jaramillo J (2012) From forest to plantation? Obscure articles reveal alternative host plants for the coffee berry borer, *Hypothenemus hampei* (Coleoptera: Curculionidae). *Biol J Linn Soc* 107: 86–94.

89. Dunning Hotopp JC (2011) Horizontal gene transfer between bacteria and animals. *Trends Genet* 27: 157–163.

90. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863–864.

91. Joshi N. Sickle - A windowed adaptive trimming for fastq files using quality. Available: <https://github.com/najoshi/sickle>. Accessed 2013 May 30.

92. Chevreux B, Pfisterer T, Drescher B, Diesel AJ, Müller WEG, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14: 1147–1159.

93. Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 18: 821–829.

94. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.

95. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

96. The UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* 41: D43–D47.

97. Zhao QY, Wang Y, Kong YM, Luo D, Li X, et al. (2011) Optimizing *de novo* transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* 12 Suppl 14: S2.

98. Li R, Yu C, Li Y, Lam TW, Yiu SM, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966–1967.

99. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, et al. (2009) *De novo* transcriptome assembly with ABySS. *Bioinformatics* 25: 2872–2877.

100. Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25.

101. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* 5: 621–628.

102. Ramsköld D, Wang ET, Burge CB, Sandberg R (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 5: e1000598.

103. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30: 772–780.

104. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690.

105. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406–425.

106. Felsenstein J (2013) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

107. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783–791.